Seybold San Francisco '93, Part II

Electronic Document Delivery and Retrieval

IN CONTRAST WITH last year, when the electronic document delivery market seemed suddenly to explode with new possibilities and lots of confusion, this year was one of technology maturation and gradual customer implementations that are providing a body of reference from which others can learn. However, despite some advancements, the market is still confusing.

In the past 12 months, page-turning technology for easily transmitting single documents electronically has gone from the preview stage to shrink-wrapped, low-cost commercial products available at retail computer outlets. In the field, page-turners such as Adobe Acrobat and No Hands Common Ground are providing real cost and productivity benefits, particularly for print-on-demand applications.

At the same time that distribution of single documents on an *ad-hoc* basis is becoming a commodity technology, delivery and retrieval of documents from *electronic libraries* is becoming more of a puzzle. Now that it is becoming easier and easier to archive compound documents, along with text, in an electronic repository, how do we go about organizing that repository? In what data formats should the documents be saved? What kinds of tools will users need to locate documents in a collection? Are those tools different for CD-ROM and so-called media servers? How should documents be presented to users for browsing or retrieval? If I'm a commercial publisher, how do I charge for accessing documents? How do I maintain copyright protection?

The answers to these and myriad of other related questions are changing so quickly that to make a cohesive picture turns out to be more like building an ongoing collage than painting a landscape. We understand the difficulty that poses when reading the pieces that follow on their own, without seeing the rest of the pieces needed to get a complete picture. But to write up what is happening in this market from one viewpoint, as if it could be told completely as a cohesive story, would not be a true reflection of what is going on. It may be completely unnerving, but for now it is best just to try to stay current with the events; the overall picture is going to take time to emerge.

Issues and Trends

From searching to knowledge extraction

As it becomes easier to file documents electronically, it is becoming harder to locate relevant information in the archives. As in a conventional library, an electronic library can catalog its items according to classification schemes and indexes. But one problem of paper filing systems and ordinary libraries also applies to electronic repositories—once the archive becomes large, it becomes very difficult to locate information by topics that have not been indexed ahead of time.

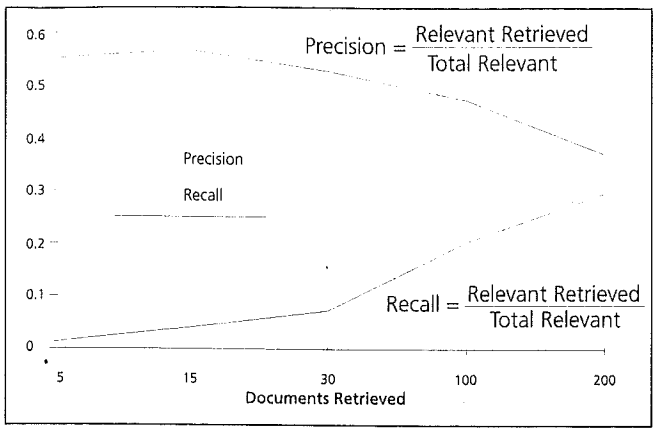
Is there any way to do a better job of extracting knowledge from electronic archives than paper ones? This question was posed by moderator Carl Frappoalo at the beginning of the session "Extracting Knowledge from Document Archives." By the end of the session, the answer was a tentative yes, but not to the degree that people would like or expect.

The panel was made up of three vendors of text-retrieval engines—ConQuest, Fulcrum and Personal Library Software (PLS)—along with John Dawes of Adobe, representing Acrobat and Verity.

Improving upon boolean logic. The vendors all discussed ways that they have improved the ease of making queries, the precision of those queries and the flexibility in running queries against multiple kinds of information. Vendors of both CD-ROM and server-based products are adding a variety of querying tools beyond conventional boolean queries. These methods include:

1. *Relevance ranking.* Most vendors now offer some means of sorting the hit list that results from a query according to how relevant the system thinks the material is to the query. In a simplest sense, this involves sorting by the greatest number of hits within a document. Some vendors apply a statistical model to derive relevance. A more advanced feature is to show a graphical representation of the clustering of the hits, which immediately gives the user feedback as to what sections or documents are particularly relevant.
2. *Lexical help.* Many vendors also extend queries with the use of word stems, thesauruses and dictionaries. With these tools, a simple query can be extended to include other related words.
3. *Concept searches.* Some systems let you group words into user-defined concepts or topics. For example, you might want a topic query, called NAFTA, to look for documents that contain Mexico or Canada within a paragraph of trade or tariff. ConQuest provides thousands of concepts out of the box; others, such as Verity, provide a way for users to create their own topics.
4. *Natural-language queries.* Because boolean queries are hard to write, some vendors have developed ways to write queries in English or other languages that people speak, rather than asking people to learn the language of the computer. The system then takes care of translating the natural language query into the language of the software.
5. *Query by example.* A quantum leap forward in full-text retrieval is the notion of querying by example. In this method, the user does not formulate a query at all. All you have to do is swipe a selection of text and tell the system to "find me more documents like this."
6. *Mixing of full-text with fielded searches.* Fulcrum has built into its latest product the concept of writing a single query that runs against both an SQL, fielded database and a full-text repository. This mixture of full-text and SQL queries represents the first step in what will undoubtedly be complex APIs for running queries against all kinds of information, text, image, video, financial, and so forth.

Despite these advances, full-text retrieval has severe limitations for finding information when you are not sure what documents



Algorithms aren't enough. Matt Koll of PLS presented this graph, which shows that with full-text retrieval systems, the more documents your query locates the lower the percentage of hits that are actually relevant to the search. His point was that for real advances in retrieval to happen, vendors will have to focus on human interaction and aspects of retrieval other than search algorithms.

you are hunting for. The information repository may contain relevant items, but the *best* full-text retrieval can do in most cases is find 25% of the relevant information. (ConQuest claimed that it could get that number closer to 50%; Matthew Koll of PLS, which has had much more experience with very large archives, disputed that claim. We'd note that even 50% is not great, especially if the crucial documents are in the half not found.)

The upshot is that full-text retrieval *is* a big improvement over a pure keyword/topic/author/title index, such as you get in the library. Because of full-text retrieval, you can locate information in the repository that would otherwise be overlooked. But because we as humans tend to change our questions (sent to the system as queries) as we start to get answers (looking at the hit list and sampling documents), the percentage of recall and precision never looks great, because we are giving the system a moving, rather than fixed, target.

Navigational aids. One aspect the speakers missed that was very evident on the show floor is the use of navigational aids to supplement full-text retrieval. In systems designed for delivering or housing large information collections, navigational tools are proving to be a useful supplement to retrieval by queries.

1. *Logical collections.* The first, most obvious idea is to group similar documents together so that, for example, going to the electronic file cabinet of monthly sales reports and opening the drawer for those from the Western sales region shows you the Western sales reports. This idea can be extended to topical collections. SuperBook uses the metaphor of a bookcase and bookshelves, where each shelf might be a subject, similar to the way libraries organize their shelves.
2. *Structural navigation.* It is all well and good to locate documents that contain words or phrases of interest, but for large documents you also want some notion of where you are in the document and, at the same time, you want a tool for navigating according to the contents, rather than simply jumping to the next hit or scrolling through a humongous text file. SuperBook is a useful example here, too. The product pioneered the concept of showing the user a collapsible/expandable outline, with full-text hits shown against each element of the outline. EBT also uses this technique in DynaText. Many other vendors provide structural navigation without the link to full-text hits.
3. *Random hyperlinks.* Ted Nelson pioneered the notion of nonlinear navigation, or hypertext, nearly 20 years ago. Today, because the links often go to information other than just text, we may call them hyperlinks, but the idea is still the same—to create the electronic equivalent of cross-references in paper. Such links may be to material that is not found in a full-text search, or it may be outside of the domain of the text-retrieval system (as is currently the case with images and video). In electronic repositories, it can be much easier to follow links, because the software will take the user directly to the referenced item, rather than asking users to pursue those cross-references on their own.
4. *Directed paths.* Random links are OK for simple cross-references, but typically the user has no idea where the links are in the repository ahead of time. With a product such as Acrobat, for example, you can have any number of pages linked to each other, but there is no way to extract a series of links into a path that someone else might want to follow. The notion of directed paths, as seen in a product such as Westinghouse Pathways, is one we think will prove very useful for large document collections. It will first be used by editors who have control over a collection. For example, an editor of documentation manuals might want to establish different paths for diagnosing and correcting a problem according to the skill level of the reader. The editor of an electronic encyclopedia might want to establish a path that relates to a subject, such as jet propulsion, in which a story is told through a series of articles and screens that may or may not have the word jet propulsion in them.

Everybody offers everything. Once you move beyond electronic delivery of single documents (that is past products like No Hands Common Ground and Farallon Replica) there are few easy distinctions to draw among the access methods of different products. Some or all of these retrieval and navigational methods may be present in any product that handles retrieval of multiple documents. As electronic libraries become more prevalent, and the modularity of electronic access increases, we will undoubtedly see ways to combine all sorts of retrieval tools with the leading archive formats.

There is no longer a clear market distinction between tools aimed at commercial publishers and those designed for inhouse use. At one time, CD-ROM authoring tools were aimed primarily at commercial publishers. Today half of the CD-ROMs produced are created by inhouse publishers, and it is becoming increasingly easy for them to make these discs directly from PostScript, SGML and several page makeup formats, as well as with the more advanced tools for CD authoring. With the advent of digital highways, server-based products such as SuperBook or OracleBook become viable for commercial publishing as well as the inhouse use from which they originated. Obviously, as we'll discuss in a moment, the copyright and intellectual property concerns of commercial publishers are often different from those of inhouse publishers. But people on both sides will be using similar tools for delivering and retrieving information.

Maps of the paths to knowledge. As we gain more experience in electronic archives, it is quite possible that *users* will begin to make directed paths. In the same conference session, the full-text vendors argued that saving queries is not very helpful because it is too hard to say who the expert query-makers should be. And if everyone saves their queries, before long there are too many for users to know which to choose. We agree that just saving queries randomly in a pile is not helpful.

But we disagree that past experience has no relevance to future searches. Surely the feedback from those crossing the Rocky Mountains proved useful to those who followed in their path. If you were looking for a good place to mine or farm, then feedback such as that path leads to a vast desert could save you a lot of wasted time and energy. This is just as true today, when we ask a librarian

where to begin looking for information about a topic in the vast repositories that the library offers access to. If you want to know where to look in an electronic archive for an explanation of cost accounting, you might begin by asking the librarian what path to follow, rather than expecting some presaved query on cost-accounting to be the most relevant.

There are often several paths one can take in looking for information, just as there are many ways to travel and many roads to drive on. When we want to know how to drive from Cincinnati to Columbus we can ask directions or buy a map. You might take the scenic route; maybe you choose the interstate highway; the map indicates that not all roads (hyperlinks) are the same.

Similarly, we expect that within an information collection, there are certain reference points that will be used more often than others. Obviously the major reference points, just like cities in our travel analogy, will be destinations for many more paths than the obscure titles. We can imagine that if visual navigation aids were available, they might help us figure out in what way we intend to investigate the electronic library.

To make this work, users will need tools for naming, saving, annotating and visually representing paths through the collection. As we mentioned, these tools will arrive first for authors and editors. We expect that the feedback they give vendors should be useful in developing tools for more general-purpose use.

Eventually, with navigation methods such as these, combined with full-text retrieval and (perhaps most important of all) tools for user interaction, we may one day have an interface for extracting knowledge from electronic archives.

Digital highways

High-bandwidth telecommunication is changing both the production and the distribution of information—and the rate of change is accelerating with the continual introduction of fiber optics, digital phone switches and so forth. Some of the changes simply make ordinary data transfers go faster. In 1976, high-speed data communication meant a 300-bps modem. Since then, modems have gotten faster; the latest models (using V.FAST protocols) can go up to 38 kbps. But that's about as far as tone-based signaling can be pushed. To go faster, it is necessary to turn to pure digital techniques. For this reason, many of the booths on the Expo floor were using ISDN phone lines (furnished by Pacific Bell) this year, where a year ago they would have been using modems and voice-grade lines.

There are a lot of good analogies between phone lines and roads, starting with the fact that both are communication systems in the broadest sense. Both are costly to build and maintain, but are taken for granted by their users, at least until something goes wrong. Other parallels include:

- *Failures.* It pays to build a redundant network, with more than one way to connect any two points, as a defense against inevitable breakdowns.
- *Bottlenecks.* As soon as you try to ease a bottleneck by upgrading the weakest component of the system, the bottleneck moves to the next-weakest part. In a road system, the bottleneck often occurs because the on-ramps and access roads aren't as well engineered as the freeway. Digitally, the same problem occurs at the interface between the high-speed backbone network and the slower local loops.
- *Peak loads.* There are two ways to cope with peak demand: design the system to handle the peaks (which means high costs and complexity, along with underutilization most of the time) or flatten the peaks by making the users queue up for service (which leads to traffic jams, complaints and failure to make deadlines).
- *Traffic jams.* The availability of bandwidth generates new uses for it, so that any channel eventually fills up.

Restructuring production. Brian Anderson, of Contex Systems, described how a newspaper group based in Sydney, Australia, uses ISDN to move data from the editorial offices on one side of town to the printing plant on the other side. The link between the two sites was painless to create: buy the boxes (a network router and a line interface at each end), order lines from the phone company, plug everything in and pay the monthly line fees.

In operation, the link is quite transparent. At each site, the file servers at the other end of the line behave the same as the local servers, albeit somewhat slower. Nor is there a difference in access techniques or user interface. Users running Windows or Macs send their files across town by simply dragging icons.

The cost is less than sending disks across town by couriers on motorbikes: about $9 per day for dedicated lines up to 25 kilometers long and rated at 256 kbps.

The lines have enabled a significant change in production workflows. In the old days, a publisher would deal with a service bureau for outputting films, a trade separator for scanning color images and a printer. Now, the scanning is done inhouse with desktop drum scanners and fully made-up pages are transmitted as PostScript files to the printer, who outputs final films for the press. Two trade suppliers are no longer needed.

Extinction or opportunity? Speaking on behalf of Business Link, a New York City service bureau, Todd Melet opined that digital highways need not mean bankruptcy for trade specialists. Rather, they might have the opposite effect of opening new markets, allowing the specialist to serve customers at much greater distances than before. For example, Business Link uses ISDN and Switch-56 data links to serve roughly a thousand ad agencies and designers. Most are in New York, but he has some customers clear across the country.

Melet noted that, while a daily newspaper could easily justify a dedicated line between its editorial and printing sites, most design shops don't do enough business with any one supplier to justify the equipment cost. However, ISDN and Switch-56 are dial-up services; you can call up any properly equipped phone whenever you need to. The trick is to develop applications that go beyond simply sending pages. Ordering images from a stock-photo dealer, placing display ads in newspapers, interactive design conferences, interconnecting LANs and even telecommuting from an at-home office are all possible now. At today's prices—$40/month in New York City—it doesn't take very many such uses to tip the scale in favor of ISDN now.

WAIS on the Internet. Publishers who are looking for an electronic outlet for their information can use an existing distribution medium: the Internet. This is a web of about 50,000 commercial, governmental and academic networks, encompassing 1.7 million host computers in 91 countries. It used to be restricted to non-commercial traffic, a legacy of its origins as an academic research project funded by the U.S. government. Now, however, commercial traffic (carried over non-government-funded wires) is the fastest-growing component of the Internet.

To help publishers seeking electronic outlets, WAIS (Wide Area Information Servers) was founded as a for-profit corporation *(see Digital Media, Vol. 1, No. 9)*. It has defined standard server protocols for searching text and image databases, and it offers server technology and access tools to its customers.

The result: An information provider now can have a global market, reaching many more potential customers than would be possible with a private service. In his conference presentation, WAIS director John Duhring described the experience of Counterpoint Publishing, which handles the online version of *Commerce Business Daily*. Initially, Counterpoint had a line-oriented user interface and direct dial-up access to its server, and typically transmitted about 2,000 documents per month to subscribers. It then adopted the Gopher user interface (developed at the University of Minnesota, home of the

Alphabet Soup, Telco Style

Believe it or not, the telephony business has even more acronyms than the computer industry. ISDN stands for integrated services digital network, and for practical purposes, it means one twisted-pair circuit carrying data at 64K bits per second. You can gang multiple circuits together to obtain higher speeds. By digitizing the sound, one circuit is sufficient for two separate audio conversations plus some 9,600-bps data on the side. This is how it is being marketed for at-home offices, for example.

ATM, in this business, has nothing to do with cash machines. It stands for asynchronous transfer mode, which is going to be the next big communication and wide-area networking standard. It runs at a wide range of speeds, but it is cost-effective now only for greater-than-10-megabit/second lines.

There is also a phone-company acronym for an ordinary rotary-dial voice line. It is POTS, which stands for plain-old telephone service.

Fighting Gophers) and joined the Internet; now a typical month sees about 50,000 documents transmitted.

There are some restrictions on the way business may be transacted over the net. These are mainly due to Internet "culture" rather than to laws. The Internet is not a single entity, but a loose consortium of cooperating entities, all of which own and fund their piece of the web. It is governed by an ethos of courtesy and reciprocity, not by laws and regulations. Thus, Rule One says that all network services shall be passive; someone must request a document before it can be transmitted. There shall be no "junk E-mail" of advertising broadsides to solicit business.

There are also some issues of copyright infringement. The belief that all information ought to be free still lingers in many corners of the Internet, and many an information provider has been appalled to find his stock in trade being openly posted on public-access bulletin boards. But in fact, the danger is not much different from what a CD-ROM publisher would face—once the discs are out there, the database is exposed. And copyright law still applies to Internet data, just as it does to CD-ROMs. As a practical matter, said Duhring, copiers may be stealing your intellectual property, but they are also advertising for you.

WAIS likes to use a retail-store metaphor for its information servers. The customer can come into the store and look around, browsing through the available topics. Perhaps you (the merchant) will allow him to read the headlines or the abstracts; but at some point that you have selected, you can require a fee before divulging any more. Payment can be by the document (perhaps a credit card number encrypted by a public-key algorithm), by annual subscription or any other mechanism that fits your business model.

Impact on mom and pop. Steve Waters, VP of the Rome (NY) *Sentinel,* spoke about the implications of electronic delivery of information. Newspapers used to think that it would be impossible to provide a true electronic equivalent of the daily paper, but they are now changing their minds. Most of the necessary hardware—computers, phone lines, software—is available now. The only missing piece is a large, high-resolution screen, and that is likely to arrive soon. (Knight-Ridder's Roger Fiddler estimates that in 7–10 years you'll have a tabloid-size screen costing $200. It will be half an inch thick and you will be able to take it into the bathroom.) The other issue is the publisher's mindset, but publishers are starting to realize that a newspaper is just a format. An electronic newspaper can also be formatted with type, ads, headlines and the other navigational landmarks.

Collecting local information and editing it into stories is what newspapers are really all about. Writers may have to learn to write dual news streams—one for print, the other for the screen—but no other agency is fitted to perform this task. Other newspaper departments will have similar maturations; the layout staff, for example, will become adept at placing hypertext buttons on the screen. Advertising departments will take on the media-selection functions now done by ad agencies. And the business departments will learn how to sell computing and information services.

Electronic information will raise other problems. One will be copyright law, which is far behind the pace of innovation. The law must be updated to recognize separate uses for information, including referential use (hypertext pointers), reportorial use (now covered by the "fair use" clause), advertorial use (commercial quotation), anthological use (full-text inclusion) and artistic use.

Another social problem may be tougher. Hometown retail shopping has been the economic *raison d'etre* for small towns for nearly a century. Now, catalogs and TV shopping networks are gradually eliminating the local mom-and-pop stores and thus threatening the existence of the towns. Perhaps the stores will be transformed into cottage-industry fulfillment operations, executing the orders generated by the advertising in the new electronic medium.

Fortunately, Waters said, publishers don't have to predict the future very far or very accurately. They just have to be sure that they don't dead-end in the meantime.

Guarding your intellectual property

One of the reasons for adding an electronic outlet for your publications is to provide richer forms of content: sound, animation and movies. But as soon as you venture into these art forms, you are no longer in the familiar world of print, and different laws now govern your rights and duties. The legal structures of Tin Pan Alley and Hollywood, not Fleet Street, govern the new media.

Along with the programmers, audio technicians, animators and other skilled personnel, you should be sure to have a competent intellectual-property lawyer as an early member of your team. His mission: to make sure that you actually have the legal rights to use the sounds and images that you want to use in your multimedia publication. As we shall see, that's not so simple a task.

It starts with copyright. Most of the rights you need have their basis in the legal concept of copyright. This includes the right to reproduce and distribute a work, to modify it or make derivative works from it, and to exhibit or perform the work in public. There are also the well-known laws of patents and trademarks.

However, all of these laws vary from country to country. For example, copyright generally lasts for the life of the creator plus 50 years. But in Germany it lasts for life plus 70 years. Before the Second World War, copyright in several nations lasted only for 20 years unless renewed, and thus many movies produced before the Great Depression fell into the public domain. But this must be verified on a case-by-case basis.

There is also an entirely separate class of personal rights. These include:

- *Privacy.* The use of people's images and voices for commercial purposes is a right that must be secured from the individuals themselves. The privileges that apply to news coverage, which essentially strip away the privacy rights of politicians and celebrities, typically don't extend to productions for profit.
- *Publicity.* People who have built up a public reputation have the right to control how others exploit that reputation. For exam-